基于深度学习的人体动作识别方法 *

李玉鹏，刘婷婷，张　良

(中国民航大学 天津市智能信号与图像处理重点实验室，天津 300300)

摘　要：针对人体动作深度视频的四维信息映射到二维空间后，动作分类容易发生混淆的问题，提出一种基于深度学习的人体动作识别方法。首先构建空间结构动态深度图，将深度视频的四维信息映射到二维空间，进行信息降维处理；然后提出基于联合代价函数的深度卷积神经网络，结合交叉熵损失函数与中心损失函数作为联合代价函数，指导卷积层学习到更具分辨力的深度特征，以进行更精确的分类。在 MSRDailyActivity3D 和 SYSU 3D HOI 两个数据集的实验结果表明，与现有方法相比，该方法识别率得到了较明显地提升，验证了该方法的有效性和鲁棒性。该方法较好地解决了动作分类容易发生混淆的问题。

关键词：深度信息；人体动作识别；深度学习；空间结构动态深度图；深度卷积神经网络

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2018.05.0499

Human action recognition based on deep learning

Li Yupeng, Liu Tingting, Zhang Liang

(*Tianjin Key Laboratory of Advanced Signal & Image Processing*, *Civil Aviation University of China*, *Tianjin* 300300, *China*)

Abstract: In order to solves the problem of action classification prone to confusion after mapping the four-dimensional depth information to two-dimensional space, this paper proposed a method for human action recognition based on deep learning. Firstly, the method constructed spatially structured dynamic depth images for dimension reduction. Then, it proposed the deep convolution neural network with joint cost function, which combined the cross entropy loss function and the central loss function as cost function, to guide the convolution layer to learn more discriminative deep features. The experimental results evaluated on the public MSRDailyActivity3D dataset and SYSU 3D HOI dataset. It show that the method obtain a better performance compare with other existing method, which validate the effectiveness and robustness of the method. The method effectively solves the problem of action classification prone to confusion.

Key words: depth information; human action recognition; deep learning; spatially structured dynamic depth images; deep convolution neural network

0　引言

人体动作识别在智能监控、人机交互、视频检索、虚拟现实等方面具有广泛的应用，因此其一直是计算机视觉领域一个活跃的研究方向。在以前的研究中，很多关于人体动作识别的研究方法都集中在传统的 RGB 视频[1-4]，但基于 RGB 视频数据的处理有很多难点，比如：不具有视角不变性，对光照和背景的变化敏感，对噪声不鲁棒等，虽然近几年通过研究者的努力，取得了一些很有意义的成果，但人体动作识别的研究仍然非常具有挑战性。

近年来，微软 Kinect 的发布为这一领域带来了新的机遇，Kinect 设备可以实时地采集深度图，与传统彩色图像相比，深度图有许多优点，例如，深度图序列实质上是四维空间可以包含更丰富的动作信息，对光照条件的变化不敏感，可以更可靠地估计人体轮廓和骨骼等[5]。文献[6~11]利用深度图的这些特性设计出专门的特征描述子，一定程度上对动作识别领域产生了深远的影响。Liu 等人[12]提出增强的骨骼点形象化方法利用骨骼点的时空序列对人体动作进行视角不变的识别，具有更广泛的实用性，但仍受限于利用骨骼数据构造特定的特征。因此，上述方法都是基于手工制作的特征，这些特征是对局部或全局时空信息的浅层次描述，无法同时捕获动作中重要的时空和结构信息。

随着深度卷积神经网络（deep convolutional neural network, DCNN）在 ImageNet 图像分类竞赛中获得巨大的成功[13]，许多研究者将 ImageNet 上训练好的模型应用到诸如属性分类[14]、图像表示[15]和语义分割[16]等任务中，取得了良好的效果。然而，

收稿日期：2018-05-03；**修回日期：**2018-07-09　**基金项目：**国家自然科学基金资助项目（61179045）；民航安全能力建设项目（20600523）

作者简介：李玉鹏（1993-），男，广东英德人，硕士研究生，主要研究方向为计算机视觉与深度学习（yupengli666@126.com）；刘婷婷（1992-），女，硕士研究生，主要研究方向为图像处理与计算机视觉；张良（1970-），男，教授，博士，主要研究方向为图像处理、模式识别、计算机视觉.

上述研究均是针对彩色图像的图像理解任务，人体动作识别不同于一般的图像理解任务，特别是基于深度信息的人体动作识别问题，其以深度视频这种四维空间的形式表示，因此无法效仿上述任务直接使用 DCNN 进行识别。

Wang 等人[17]尝试通过设计加权的深度运动映射图作为 DCNN 的输入，使人体动作识别问题转换为图像分类问题,首次使用 DCNN 对基于深度图的人体动作进行识别，但实验结果表明该方法的鲁棒性欠佳。受 Fernando 等人[18~20]提出的顺序池化法（rank pooling, RP）在基于彩色图像的人体动作识别方向上取得较大成功的激励，Wang 等人[21]在 RP 的基础上提出空间结构动态深度图（spatially structured dynamic depth images，SSDDI），克服了 RP 操作抑制深度图空间局部细粒度运动信息的缺点，达到了较高的识别率。

以上分析可知，目前的研究工作集中在寻求设计某种有效的特征表示方式，期望在动作的四维信息映射二维空间后，尽量将动作的重要特征在二维空间中得到表征，从而提高动作识别的准确率。然而，笔者在研究中发现，动作的深度信息被映射到二维空间表征后，动作在分类过程中很容易产生混淆，从而限制该类方法的识别率上限。

针对现有方法所存在的问题，通过对文献[22]的研究和实践，受 Wen 等解决人脸识别领域类似问题所采用方法的启发，笔者从神经网络提取特征与分类的机制考虑问题，结合基于深度图的人体动作识别的特点，提出基于联合代价函数的深度卷积神经网络（joint cost function based deep convolution neural network, JCF-DCNN）用于人体动作识别，尝试提高动作分类的准确性和鲁棒性，该方法在网络训练过程中增加训练样本的特征空间与类中心的距离约束，以兼顾动作特征的类内聚合与类间分离，指导深度卷积神经网络学习到具有较强分辨力的特征，以促使后续进行较精确的分类。图 1 是该方法的整体流程示意图，在 MSRDailyActivity3D 和 SYSU 3D HOI 两个数据集的实验结果表明，使用本文提出的方法，人体动作识别的准确率和鲁棒性得到了明显的提升。

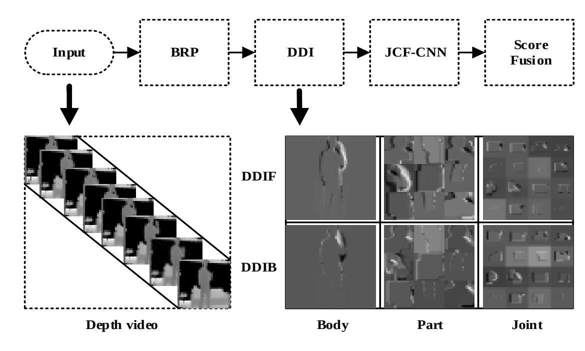


图 1　整体流程示意图

Fig.1　Overall method flow

1　动作表征方法

1.1　RP 与 BRP 的工作原理

将 k 帧的图像序列表示成 $X=<x_1,x_2,...,x_t,...,x_k>$，$\varphi(x_t)\in R^d$ 表示从每一帧 x_t 中映射而来的特征向量，特征向量 $\varphi(x_t)$ 前 t 帧的平均值用 $V_t=\frac{1}{t}\Sigma_{\tau=1}^{t}\varphi(x_t)$ 表示，对任意时序 t 指定 $r_t=w^T\bullet V_t$ 表示其得分，一般来说，时序越靠后，对应的得分会越大，因此得分函数 r_t 满足约束：$r_i>r_j\Leftrightarrow i>j$。顺序池化过程的目的是找到满足目标函数式（1）的 w^*，

$$\underset{w}{argmin}\frac{1}{2}\|w\|^2+\lambda\sum_{i>j}\varepsilon_{ij}\qquad(1)$$
$$s.t.\ w^T\bullet(V_i-V_j)\geq 1-\varepsilon_{ij},\varepsilon_{ij}\geq 0$$

其中：ε_{ij} 是一个较小非负值，参数 w^* 可以表征在 v_t 刚开始但 v_{t+1} 尚未进行之时对应的图像序列的信息，其可作为图像序列的特征描述子。

由上述分析可知，RP 是无监督学习过程，可以将图像序列描述为新的特征，其是与输入图像等尺度的二维空间。由于其包含整个动作过程时空变化的信息，因此称之为动态图（dynamic image, DI），基于深度信息的动态图则称为动态深度图（dynamic depth images，DDI）。

由于在进行 RP 时，截止到时间 t 的平均特征 v_t 被用于对帧 t 进行分类，所以经池化后的特征偏向于图像序列的起始帧，导致起始帧对于 w^* 的影响更大。然而，这在动作识别中显然是不合理的，因为并没有先验知识可以得知哪一帧对该任务更重要。

双向顺序池化法(bidirectional rank pooling, BRP)可以大幅度的降低上述偏差。如果将上文所提及的过程称为正向 DDI(Forward DDI, DDIF)的生成过程，则将图像序列反向排序后再进行 RP 即为反向 DDI（backward DDI,DDIB）的生成过程，同时产生 DDIF 和 DDIB 的方法即为 BRP，由此，每个动作的深度图像序列经 BRP 后最终将生成 DDIF 和 DDIB 一对图像。

C1	C2	C3
C4	C5	C6
C7	C8	C9

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

图 2　SSDDI 组件分布

Fig.2　SSDDI Component distribution

1.2　SSDDI

文献[18,21]的研究表明，BRP 不仅受限于动作长期的动态过程，而且也受限于空间域。由于非监督的学习方式，BRP 在时域中主要对突出的全局特征进行编码，却没有同时在时空域发掘出具有分辨力的运动模式。因此，若直接使用 BRP 对动作进行处理，将导致空间中颗粒度较小，但却对动作识别具有较高区分度的运动信息被颗粒度较大的运动信息所抑制，特别是对细粒度的动作来说，在整个动作过程中，其局部的时空子空间运动信息相比于全局的运动信息来说更重要。SSDDI 将深度图像序列在空间域中按不同的颗粒度分解为多个部分，再对各

部分分别进行 BRP 操作，最后其组合起来作为新的表征，可以有效的解决上述问题。

具体来说，将深度动作序列提取前景后，根据骨骼点数据作为引导，在空间域分别按照全身区域（body）、部分区域(part)、骨骼区域(joint)三个层次分解；其中 body 层次是将包含 20 个骨骼点的整个人体，由于只有一个组件，所以其进行 BRP 后形成的 DDI 即为 SSDDI；part 层次的组件构成如表 1 所示，每个组件的区域由 3 个骨骼点之间的最大距离确定，共分成 9 个部分，作为 9 个组件，可以覆盖全身，将每个组件分别 BRP 后生成对应的 DDI，按照图 2 左侧所示构造成 SSDDI；joint 层次的每个组件包含 1 个骨骼点，从表 2 可以看到，该层次的 SSDDI 共有 16 个组件，每个组件的区域由骨骼点所在位置对外拓展一定的距离而成，用于组件的骨骼点是从全部 20 个骨骼点中选取的噪声较低的 16 个，组件的分布见图 2 右侧。

表 1　Part 层次各个组件包含的骨骼点

Table 1　Components of part level

C1	head, shoulder center, shoulder left
C2	head, shoulder center, shoulder right
C3	elbow left, wrist left, hand left
C4	elbow right, wrist right, hand right
C5	spine, hip center, hip right
C6	spine, hip center, hip left
C7	knee left, ankle left, foot left
C8	knee right, ankle right, foot right
C9	shoulder left, shoulder center, shoulder right

图 1 中展示了同一动作对应的三个层次的 SSDDI，从中可以看出，相比于 body 层次的 SSDDI，part 和 joint 层次的 SSDDI 对颗粒度较小的动作表征更具有分辨力，可以更有效的实现对动作进行从全局到局部运动以及结构信息的表征，将上述三个层次的 SSDDI 分别训练 JCF-DCNN 后进行决策层融合，有利于提高动作识别的准确率。

表 2　joint 层次各个组件包含的骨骼点

Table 2　Components of joint level

hip center	spine	shoulder center	head
shoulder left	elbow left	hand left	shoulder right
elbow right	hand right	hand left	knee left
foot left	hip right	knee right	foot right

2　JCF-DCNN

2.1　JCF-DCNN 的网络结构及超参数设置

JCF-DCNN 的意义在于其具有较强的特征学习和分类能力，可以提高对 SSDDI 这类图像样本的分类准确率，换句话说，JCF-DCNN 可以将两个相似度较高但属于不同类别的样本区分开来，以对其进行正确归类，降低动作分类的混淆程度。

其主要的特点在于联合交叉熵损失函数及中心损失函数作为网络分类层的代价函数，也即在网络训练过程中增加样本特征空间与类中心的距离约束，这样可以指导 JCF-DCNN 的卷积层在训练时可以学习到更具分辨力的特征。

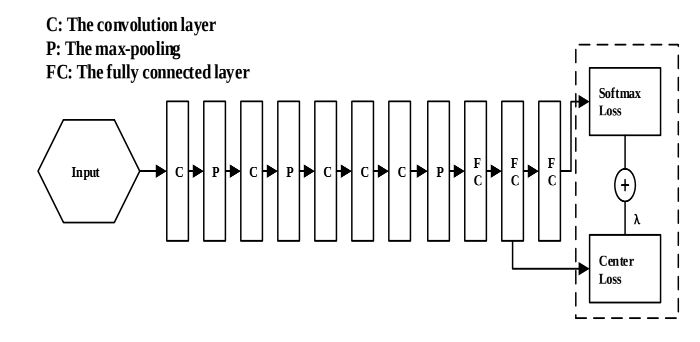


图 3　JCF-DCNN 网络结构

Fig.3　JCF_DCNN structure

如图 3 所示是所提出网络的具体结构图，可以看出，JCF-DCNN 具有 12 层，主要包含 5 个卷积层和 3 个全连接层以及后端的分类层，该网络分类层的联合代价函数由交叉熵损失函数与中心损失函数组成，表 3 记录了该网络架构卷积层和前 2 个全连接层的相关超参数设置，第 3 个全连接层网络的神经元数量与相应数据库的样本类别数一致。由于目前基于深度图的动作识别数据集规模普遍都比较小，若用其从头训练具有百万级训练参数的深度卷积神经网络，会出现过拟合现象，因此，本文采用迁移学习的方法，将已在大规模数据集 ImageNet 上预训练完成的参数，用于对本网络的全部卷积层和前 2 个全连接层进行参数初始化。

训练时，按上述方法初始化相应的网络层，但最后一个全连接层使用均值为 0、标准差为 0.01 的高斯分布随机初始化。经过多次实验分析比较，将前 3k 次迭代学习率设置为 0.001，后续的 3k 次迭代学习率设为 0.0001，共迭代训练 6000 次后能够达到良好效果。动量和权重衰减因子使用经验值 0.9 和 0.0005。为进一步避免过拟合，在图像样本进入深度卷积神经网络前，先对其尺度变换至 256×256，然后以中心和四角为坐标原点，剪裁出 224×224 的区域，再进行镜像操作，使实际的训练样本达到输入样本量的 10 倍。但测试时只截取测试图像样本的中心区域，且不进行镜像操作，同时由于测试阶段只进行前馈操作，因此并不涉及中心损失函数，只需将交叉熵损失函数的输出值进行均值融合。

表 3　JCF-DCNN 网络结构超参数设置

Table 3　Super parameter setting for JCF-DCNN

Layer	C1	C2	C3	C4	C5	FC1	FC2
numb	96	256	384	384	256	4096	4096
filter	11^2	5^2	3^2	3^2	3^2		
stride	4	1	1	1	1		
pad	0	2	1	1	1		

2.2　交叉熵损失函数

深度网络中的代价函数是整个网络模型的“指挥棒”，通过样本的预测结果与真实标记产生的误差反向传播指导网络参数学习与表示学习。交叉熵损失函数是目前深度卷积神经网络中

最常用的分类损失函数，其形式为

$$L_S=-\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^T x_i+b_j}} \tag{2}$$

其中: $x_i\in R^d$ 表示属于 y_i 类的第 i 个深度特征，d 是深度特征的维数，$W\in R^{d\times n}$ 是最后一个全连接层的权值矩阵，$b\in R^n$ 是偏置项，$W_j\in R^d$ 表示权值矩阵 W 的第 j 列，m 和 n 分别是每批训练样本的数量以及相应的类别数，由于偏差项对性能的影响极其微小，因此为了简化分析，往往可以被忽略。

由式（2）可知，交叉熵损失函数具有结构简单，计算量小的优点，因而得到了广泛的应用，然而从实际应用角度来说，该损失函数仅仅关注了待识别图像应该属于哪个类别的问题，即类间分离问题，但没有考虑同样重要的一个问题，即最终的分类器决策面区域内的空间是否均应属于该类别。实际上，同一类别下两个图像样本的距离有可能比不同类的距离还大，若在这种情况下使用交叉熵损失函数作为神经网络的代价函数，极易出现由于待分类的图像样本太相似而被误判的情况。值得注意的是，交叉熵损失函数在测试阶段不再进行梯度的计算，也不进行梯度的反向传播，仅作为计算相应类别概率值的一个函数使用。

2.3　中心损失函数及联合代价函数

为弥补交叉熵损失函数存在问题，中心损失函数给每一类数据定义一个中心点，这个中心点和聚类问题中的中心点十分相似，目的是为了使同一类数据计算出来的特征都能靠近自身类别的中心点，聚合类内特征，特征离中心点越远，则对其惩罚越大。式（3）中对中心损失函数作了形式化的表征。

$$L_C=\frac{1}{2}\sum_{i=1}^{m}\left\|x_i-c_{y_i}\right\|_2^2 \tag{3}$$

$$\frac{\partial L_C}{\partial x_i}=x_i-c_{y_i} \tag{4}$$

$$\Delta c_{y_i}=\frac{\alpha}{2}\sum_{i=1}^{m}(c_{y_i}-x_i) \tag{5}$$

其中：$c_{y_i}\in R^d$ 可由式（4）（5）计算得出，表示 y_i 类深度特征 x_i 的类中心，其随着深度特征 x_i 的变化而更新，参数 α 是范围在 $[0,1]$ 的缩放因子，通过调节此参数，可以对神经网络进行进一步的优化，因此其是一个超参数，变量 m 与 x_i 含义与交叉熵损失函数中介绍的一致。

为整合交叉熵损失函数和中心损失函数的优点，同时使深度特征类内聚合与类间分离，所提出的 JCF-DCNN 采用联合交叉熵损失函数与中心损失函数作为该网络分类层的代价函数：

$$\begin{aligned}L&=L_S+\lambda L_C\\&=-\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T x_i+b_{y_i}}}{\sum_{j=1}^{n}e^{W_j^T x_i+b_j}}+\frac{\lambda}{2}\sum_{i=1}^{m}\left\|x_i-c_{y_i}\right\|_2^2\end{aligned} \tag{6}$$

λ 是为了控制两个损失函数的比例而引入的超参数，当 $\lambda=0$ 时，联合代价函数将退化成交叉熵损失函数，图 3 虚线部分指示了联合代价函数在 JCF-DCNN 中所处的具体位置。

表 4　MSRDailyActivity3D 数据集中的识别率对比

Table 4　Comparison of recognition rates in MSRDailyActivity3D

方法	识别率
IPM [24]	83.30%
WHDMMs+ConvNets [17]	85.00%
SNV [9]	86.25%
DS+DCP+DDP+JOULE-SVM[23]	95.00%
Range Sample [10]	95.63%
MFSK+BoVW [25]	95.70%
SSDDI[21]	97.50%
本文方法	99.38%

3　实验与结果分析

3.1　实验环境与方法

实验环境所用 GPU 为 NVIDIA Quadro P2000，操作系统为 Ubuntu14.04，配置并编译 caffe-HAR(https://github.com/liyupeng-ing/caffe-HAR)，caffe-HAR 是考虑到便于读者复现和验证本文所提出的方法而上传至开源网站的文件夹，其包含按照 2.2 节所叙述的 JCF-DCNN 网络模型文件以及相关的应用程序，其是在深度学习框架 caffe[27]的基础上，通过设计具有联合代价函数的分类层模块拓展而来。联合代价函数的超参数均按照经验值，将 α 设置为 0.5，λ 设置为 0.003。采用的 MSRDailyActivity3D 和 SYSU 3D HOI 数据集中绝大部分动作都涉及人与物的交互过程，具有较大的挑战性，上述数据集均包含彩色视频和深度视频，以及对应的骨骼点数据，本文只使用了深度图像和骨骼点数据，并没有用到数据集中的彩色图像。

实验过程包含训练阶段和测试阶段。训练阶段，使用训练样本的 body、part 和 joint 这三个层次的 SSDDI 分别训练 3 个网络模型；测试阶段，用三个层次测试样本的 SSDDI 分别输入上述对应的网络模型，对每个网络模型的分类层的输出结果进行融合，取融合后得分最大值对应的标签为识别结果，实验中使用的融合方法为均值融合，测试阶段分类层中仅有交叉熵损失函数进行工作，并不涉及中心损失函数，即此时的 λ 将自动设置为零，此时分类层的联合代价函数退化为交叉熵损失函数。需要注意的是，每个 SSDDI 都对应有 DDIF 和 DDIB 一对图像，因此需要先进行各个层次内的融合，再进行层次间的融合。具体来说，body 层次的 SSDDI 对应的 DDIF 和 DDIB 输入网络模型后会输出 2 个相对应的结果，需将这两个结果进行均值融合作为该层次内的输出结果，part 与 joint 同理，最后将三个层次的输出结果再次进行均值融合，作为最终结果。

表 5　联合代价函数对网络性能的影响

Table5 Influence of joint cost function on network performance

方法	Body DDI	Part DDI	Joint DDI	fusion
JCF-DCNN*	62.50%	92.50%	93.13%	96.88%
JCF-DCNN	65.63%	90.63%	93.75%	99.38%

3.2　识别结果与分析

MSRDailyActivity3D 数据集由 Kinect 深度摄像机采集，包含 16 种动作，由 10 人完成，每人分别做 2 次动作，其中一个站立完成，另一个坐着完成，共有 320 个文件。为公平起见，训练样本和测试样本的选取均遵从文献[7]，将 2，4，6，8，10 号表演者的动作用于训练，1，3，5，7，9 号表演者的动作用于测试。表 4 是各种方法的比较结果，可以看出，基于 JCF-DCNN 的人体动作识别方法的准确率达到了 99.38%，比 SSDDI 提高了 1.88%。

为体现 JCF-DCNN 中采用的联合代价函数在该网络中所起的作用，本文设置了去除中心损失函数的 JCF-DCNN *实验，如表 5 所示，除 part 层次的融合结果较低外，JCF-DCNN 的其他层次以及最终结果均高于 JCF-DCNN *，而导致 part 层次中融合结果较低的原因可能是由于 part 层次的样本间相似度较低，不符合 JCF-DCNN 类内聚合的特性所致；值得注意的是，JCF-DCNN 整体的识别率相对于 JCF-DCNN *来说，有大幅度的提升，提升幅度达到了惊人的 2.8%，说明了联合代价函数在网络性能的提升中起到了明显的效果。

SYSU 3D HOI 数据集共包含 480 个动作文件，由 40 个人每人演示 12 种动作通过深度摄像机采集而来。在 SYSU 3D HOI 数据集上，训练和测试样本的选取与文献[23]保持一致，如表 6 所示，本文提出的方法在该数据集的识别率到比 SSDDI 要高出 1.66%，达了 97.08%。

本文提出的方法在 2 个数据集的表现均优于现有方法，验证了该方法的有效性和鲁棒性，这主要是由于所提出的方法具有更强的特征学习能力以及分辨力，可以提高对 SSDDI 方法产生的类间差异比较小且类内差异比较大的图像样本的分类准确率。同时，这也说明了深度卷积神经网络结合传统的动作识别方法，可以优势互补，对解决人体动作识别问题具有积极意义。

表 6　SYSU 3D HOI 数据集中的识别率对比

Table 6　Comparison of recognition rates in SYSU 3D HOI

方法	识别率
HON4D [8]	79.22%
DS+DCP+DDP+MTDA [26]	84.21%
DS+DCP+DDP+JOULE-SVM [23]	84.89%
SSDDI[21]	95.42%
本文方法	97.08%

4　结束语

本文对目前人体动作识别存在的问题进行了讨论，针对 SSDDI 存在的不足，提出了基于 JCF-DCNN 的人体动作识别方法，该方法主要特点是结合交叉上损失函数与中心损失函数作为联合代价函数，具有较强的特征学习和分类能力，能够兼顾深度特征类内聚合与类间分离，有效的降低了动作分类的混淆程度。实验结果表明，与现有方法相比，采用本方法，人体动作识别的准确率和鲁棒性均得到了明显提高。在今后的工作中，将面向包含动作种类更多的大规模数据集进行研究，同时尝试设计其他深度学习模型，尝试进一步提高人体动作识别的准确率和鲁棒性。

参考文献：

[1] Aggarwal J K, Ryoo M S. Human activity analysis: a review [J]. ACM Computing Surveys, 2011, 43 (3): 1-43.

[2] 张良，鲁梦梦，姜华. 局部分布信息增强的视觉单词描述与动作识别 [J]. 电子与信息学报, 2016, 38 (3): 549-556. (Zhang Liang, Lu Mengmeng, Jiang Hua. An improved scheme of visual words description and action recognition using local enhanced distribution information [J]. Journal of Electronics & Information Technology, 2016, 38 (3): 549-556.)

[3] 王满一，宋亚玲，李玉，等. 结合区域光流特征的时序模板行为识别 [J]. 系统仿真学报, 2015, 27 (05): 1146-1151. (Wang Manyi, Song Yaling, Li Yu, *et al.* Behavior recognition combining regional optical flow features and temporal templates [J]. Journal of System Simulation, 2015, 27 (05): 1146-1151.)

[4] 秦华标，张亚宁，蔡静静. 基于复合时空特征的人体行为识别方法 [J]. 计算机辅助设计与图形学学报, 2014, 26 (8): 1320-1325. (Qin Huabiao, Zhang Yaning, Cai Jingjing. Human action recognition based on composite spatio-temporal feature [J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26 (8): 1320-1325.)

[5] Shotton J, Sharp T, Kipman A A, *et al.* Real-time human pose recognition in parts from single depth images [J]. Communications of ACM, 2013, 56 (1): 116-124.

[6] Li Wanqing, Zhang Zhengyou, Liu Zicheng. Action recognition based on a bag of 3D points [C]// Proc of Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2010: 9-14.

[7] Wang Jiang, Liu Zicheng, Wu Ying, *et al.* Mining actionlet ensemble for action recognition with depth cameras [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2012: 1290-1297.

[8] Oreifej O, Liu Zicheng. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2013: 716-723.

[9] Yang Xiaodong, Tian Yingli. Super normal vector for activity recognition using depth sequences [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 804-811.

[10] Lu Cewu, Jia Jiaya, Tang Chikeung. Range-sample depth feature for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 772-779.

[11] Presti L L, Cascia M L. 3D skeleton-based human action classification: A survey [J]. Pattern Recognition, 2016, 53 (C): 130-147.

[12] Liu Mengyuan, Liu Hong, Chen Chen. Enhanced skeleton visualization for view invariant human action recognition [J]. Pattern Recognition, 2017, 68 (8): 346-362.



[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105.

[14] Zhang Ning, Paluri M, Ranzato M, *et al.* PANDA: Pose aligned networks for deep attribute modeling [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1637-1644.

[15] Oquab M, Bottou L, Laptev I, *et al.* Learning and transferring mid-level image representations using convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1717-1724.

[16] Girshick R B, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 580-587.

[17] Wang Pichao, Li Wanqing, Gao Zhimin, *et al.* Action recognition from depth maps using deep convolutional neural networks [J]. IEEE Trans on Human-Machine Systems. Piscataway, NJ: IEEE Press, 2016, 46 (4): 498-509.

[18] Fernando B, Gavves E, Oramas M J, *et al.* Modeling video evolution for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 5378-5387.

[19] Bilen H, Fernando B, Gavves E, *et al.* Dynamic image networks for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 3034-3042.

[20] Fernando B, Anderson P, Hutter M, *et al.* Discriminative hierarchical rank pooling for activity recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1924-1932.

[21] Wang Pichao, Wang Shuang, Gao Zhimin, *et al.* Structured images for RGB-D action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 1005-1014.

[22] Wen Yandong, Zhang Kaipeng, Li Zhifeng, *et al.* A discriminative feature learning approach for deep face recognition [M]// Computer Vision. Berlin: Springer International Publishing, 2016: 499-515.

[23] Hu Jianfang, Zheng Weishi, Lai Jianhuang, *et al.* Jointly learning heterogeneous features for RGB-D activity recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 5344-5352.

[24] Zhou Yang, Ni Bingbing, Hong Richang, *et al.* Interaction part mining: A mid-level approach for fine-grained action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 3323-3331.

[25] Wan Jun, Guo Guodong, Li S Z. Explore efficient local features from RGB-D data for one-shot learning gesture recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (8): 1626-1639.

[26] Zhang Yu, Yeung D Y. Multi-task learning in heterogeneous feature spaces [C]// National Conference on Artificial Intelligence. Palo Alto, SF: AAAI Press, 2011: 574-579.

[27] Jia Yangqing, Shelhamer E, Donahue J, *et al.* Caffe: Convolutional Architecture for fast feature embedding [C]// Proc of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.